基于 LightGBM 的多因子选股交易策略研究

摘要：在金融市场迅速发展情况下，量化投资方法被广泛引入证券交易当中，通过程序化交易来实现量化选股和资金配置被证明是有效可行的，其中多因子选股方法和机器学习预测模型的实践效果受到广泛认可。本文基于公司财务基本面因子和技术面因子，从公司的盈利能力、成长能力及市场动量情况等多方面选取量化因子来构建多因子模型。通过 LightGBM、CatBoost 等机器学习模型进行预测对比分析，并利用主成分分析法构造情绪指数，控制策略的开仓时机。实践证明，基于 LightGBM 的交易策略获得了 13.86%的年化超额收益且在引入情绪指数后能够降低策略回撤至 13%左右。

关键词：量化投资；多因子选股；机器学习；情绪指数

Research on Mulit-factor Stock Selection Trading Strategy Based on LightGBM

Abstract: With the rapid development of the financial market, quantitative investment methods have been widely introduced into securities trading, and quantitative stock selection and fund allocation through program trading have proved to be effective and feasible, the practical effects of multi-factor stock selection methods and machine learning predictive models have been widely recognized. Based on the company's financial fundamentals and technical factors, this paper selects quantitative factors from the company's profitability, growth ability, market momentum and other aspects to build a multi-factor model. Machine learning models such as LightGBM and CatBoost were used for prediction and comparative analysis, and the sentiment index that constructed by PCA was used to control the opening time of the strategy. Practice has proved that the trading strategy based on LightGBM has obtained about 13.86% returns in excess of market average return and could reduce the drawdown of the strategy to about 13% after the introduction of sentiment index.

Key words: quantitative investment; mulit-factor stock selection; machine learning; sentiment index

目 录

1 引 言

1.1 背景

随着国内金融市场的日益完善，证券、基金、期货及衍生品等金融工具的投资越来越吸引人们关注。早些年居民倾向于通过银行储蓄来获得固定的利息收入，以实现资产的增值。但随着国内证券、基金市场的拓展，部分居民开始将目光转向证券和基金投资，通过个人主动购买证券或购买基金的方式参与到债券和股票的投资当中。在国内金融投资方式日益丰富的情况下，股票以其低门槛的特点受到广大投资者的喜爱，股票投资方法也从传统的基本面分析发展到 K 线等技术分析，并且随着计算机技术的发展，结合计算机科学和数学方法的量化交易策略不断涌现。

在选股过程中，传统的基本面分析立足企业自身经营状况、企业的财务状况、发展前景、未来战略等角度并结合当前市场投资环境及行业状况分析企业未来的投资前景。该类分析方法依赖于对当前市场的正确判断以及企业信息披露的准确情况，且信息披露及相关市场情况传导需要一段时间，时效性不高，并受到投资者的主观影响。技术分析则通过对 K 线、RSI 和 MACD 等技术指标来判断股票的买卖交易时机，但多数的技术指标已经被大众所熟知，仅仅依赖技术指标交易不一定就能取得较好的收益表现。以上的分析方法往往需要投资者经过一段时间的分析、选股及交易，容易错过最佳的投资时点，而通过数量方法将投资者的投资思维及交易方式表达出来，并利用程序实现自动交易的量化交易方法显得更具有优势。

因此，在量化交易策略大量运用的背景下，本文基于盈利因子、情绪因子和动量因子等多种量化因子，运用不同的机器学习模型对上市公司股票收益率进行预测，并在滚动回测框架下进行预测对比分析，得到更加稳健且收益更高的选股模型，提出一种高收益的量化选股交易策略。

1.2 文献综述

对于量化选股交易策略的研究可以分为两个方面：一是多因子选股模型的研究；二是基于机器学习的量化交易研究。

1.2.1 多因子选股模型

国外量化选股研究起步较早，Markowitz（1952）通过数理统计方法建立均值-方差模型，通过均值衡量投资组合的收益率，方差衡量投资组合的风险程度，并以此来衡量

投资组合的收益表现。Treynor（1962）和 Sharpe（1964）在 Markowitz 的投资组合理论基础上提出了资本资产定价模型（CAPM），通过线性表达式描述资产收益与风险的关系。Ross（1976）在 CAPM 基础上提出套利定价理论（APT），通过多个因子来解释资产收益，认为风险资产的收益与多个因子存在线性关系。其后 Fama 和 French（1993）提出 FF 三因子模型，认为资产的超额回报可以由市场资产组合因子、市值因子、账面市值因子来解释。Carhart（1997）在 FF 三因子模型的基础上加入动量因子，使得模型的适用范围更广。Piotroski（2000）研究发现账面市值比因子和市盈率因子与股票收益呈正相关。Novy（2013）发现利润期望因子对股票收益存在解释作用。Fama 和 French（2015）在三因子模型的基础上添加了盈利因子和投资因子，提出五因子模型，发现其较三因子模型取得了更好的表现。Stambaugh 和 Yuan（2016）在股票发行、毛盈利能力等 11 个异象的基础上构建管理因子和表现因子，发现在此基础上构建的四因子模型反映了无法被 FF 三因子模型解释的超额收益。Daniel 等（2020）通过融资因子和盈余公告漂移因子来分别反映长期定价错误和短期定价错误，并测试这些因子对回报的解释能力。Kim（2021）发现在韩国股票市场上，相比于价值和动量等因子，规模因子可以带来更高的收益，且仅做多因子投资组合并不能带来太多的优势。Ashour 等（2023）指出风格因子投资对资产价格的影响取决于投资者的情绪，投资者情绪高涨时，正风格回报可以预测未来的股票回报，但负风格回报则不能。

国内学者对多因子选股也进行了相关研究，刘毅（2013）通过市净率、ROE 增长率等八个因子构建多因子模型，并在 A 股市场实证中发现模型取得显著收益。孙守坤（2013）通过对中小板股票进行统计分析，筛选出与股票收益和经济波动相关的因子作为选股基础，发现策略组合能够有效跑赢市场获得超额收益。刘洋、夏思雨等（2016）研究发现市现率等六大类因子与股票收益率显著相关。李倩倩（2019）认为相对强弱因子和账面杠杠因子等因子能够取得超额收益。王怡宁（2022）在高频数据的基础上建立选股模型，发现高频因子较传统因子具有更多的信息，该类因子有利于提升模型预测性能。刘宇轩等（2022）引入金融周期因子，提高了模型的收益以及稳健性。

1.2.2 机器学习量化交易

随机森林、支持向量机、LightGBM 和 GAN 等机器学习和深度学习模型受到各行业的广泛关注，许多研究者将相关模型运用到股票价格预测领域当中，比如，Miller，Keith L（2013）和 Miller，HongLi（2015）发现分类树能够有效得预测因子的收益回报。Chen 等（2015）运用长短时记忆网络 LSTM 对中国股票收益进行预测，与随机预测方法相比，LSTM 的预测准确率有较大提高。Patel（2015）通过人工神经网络、支持向量

机、随机森林和朴素贝叶斯四种算法预测 CNX Nifty 指数、标普 BSE Sensex 等指数的股价运动情况，发现随机森林效果最佳而朴素贝叶斯效果最差。Chow（2018）利用波兰破产公司财务因子数据，对比 Logistic 回归、AdaBoost、人工神经网络和高斯过程等算法判断公司破产特征的准确率。Abe（2018）利用深层神经网络、浅层神经网络和支持向量机等模型预测 MSCI 日本指数月收益率，得到深层神经网络对收益率的预测准确度最高，浅层神经网络最低的结论。Ayala 等（2021）将神经网络和随机森林等模型和 MACD 等技术指标相结合，发现将技术指标加入到模型当中更有利于形成有效的交易信号。

张茹、黄晨宇等（2010）利用 LSTM 构建多因子选股模型，认为 LSTM 较基准能够取得稳定的收益。谢合亮、胡迪（2017）构建估值、规模、动量等五方面因子，将 LASSO、Elastic Net 等多因子选股模型相互比较，发现 Elastic Net 的收益效果更好。胡宸（2019）运用逻辑回归和支持向量机构建预测投资组合，证明模型能够取得一定的收益，跑赢基准指数。王伦、李路（2020）利用 GCForest 算法与随机森林、支持向量机进行比较，发现 GCForest 在股市各阶段都显著优于对比模型。黄秋丽等（2021）将递归特征消除法和 Stacking 集成模型相结合，发现其效果优于随机测试。方毅、陈煜之等（2022）对比随机森林、多层感知机等八种模型构建因子投资策略，发现动量、反转和趋势因子对股票未来收益率的影响较大，且对小市值股票更具有预测性。刘向丽等（2023）以技术指标的买入卖出信号为基础，通过 Adaboost 模型对未来股价涨跌进行预测，同时发现其多空择时策略较多头策略更优。陈怡君、李欣雨等（2024）通过盈余公告漂移等异象因子建立量化策略，发现基于 LGBM 的多因子策略较其他量化策略有显著提升并且更为稳定。

1.2.3 文献评述

综上所述，在目前提出的量化投资策略中，在多因子选股方面使用的因子数量较少，且多集中于财务基本面因子，在当前的市场情况下，基于公司财务报表披露的财务因子难以适应快速变化的证券市场，而基于日度股票数据的情绪和动量等技术面因子表现更好，同时有部分研究者开始研究日内高频因子对股票收益的预测作用。在机器学习量化方面，国内外学者将传统的逻辑回归、SVM 模型以及近些年来较多使用的 XGBoost 和 LSTM 等模型进行对比，发现这类机器学习模型能够取得跑赢大盘的收益，但 XGBoost 等基于梯度提升树的模型能够取得更好的收益，而随机森林等模型的效果则较差。有部分学者将 Bagging、Stacking 等集成模型运用到选股预测当中也取得不错的效果。此外，

如 MACD 技术指标、股指期货对冲等也被用来和机器学习量化结合，以期得到更好的收益和回撤表现。

1.3 研究内容

第一步，利用国泰安数据库和聚宽（JoinQuant）平台作为数据来源，基于财务面和技术面数据构建因子，建立因子库，并通过因子有效性分析筛选出有效因子作为最终因子池。

第二步，选取随机森林、LightGBM 和 CatBoost 等模型作为预测模型，在经过超参数调优后对比分析各模型的预测表现，选取出表现较好的单一机器学习模型。

第三步，基于选择出的单一模型进行稳健性分析，分别针对持仓周期、交易资金量、资金配置和投资风格等参数进行调整与回测，确定该模型下最佳的参数设置。

第四步，在选股策略的基础上引入情绪指数，通过主成分分析构建情绪指数来反映市场投资者情绪，通过该指数进行策略开仓、空仓的择时判断。

本文拟解决问题：基于国内证券市场实际，验证各类因子在反应股票投资收益率方面有效性，提高选股模型的性能；比较不同理论模型在多因子选股方面的优劣势，构建高效高收益的投资策略，为投资者在量化选股模型的选择提供便利；引入情绪指数，通过投资者情绪判断开仓时机，控制股市下跌风险。

2　因子筛选与数据处理

2.1　数据处理

本文各因子数据来源于国泰安数据库和聚宽（JoinQuant）提供的数据接口。研究训练与回测时间范围为 2014 年 1 月 1 日至 2023 年 12 月 31 日，选股范围为中证 800 的指数成分股，其中剔除时间期内上市不足 1 年的新股、当期停牌股票、ST 及*ST 股票。

（1）缺失值填充：由于不同的公司规模及行业其因子值存在较大差别，为填充本文部分股票数据缺失值，采用市值分组与行业分组的方法进行分组均值填充。

（2）去极值：本文数据存在许多股票交易数据，容易出现极值，为避免极值对模型预测准确度的影响，采用标准差法进行缩尾处理，确定因子值在范围$[X_{mean}-3\cdot\sigma, X_{mean}+3\cdot\sigma]$内。

（3）标准化：为消除数据间存在的量纲差异，采用 Z-Score 的方法进行数据标准化

$$Z=\frac{x-\mu}{\sqrt{\frac{i}{N}\sum_{i=1}^{N}(x_i-\mu)^2}} \tag{2.1}$$

（4）中性化：为消除不同行业及市值股票各因子值之前存在较大差距，对因子值进行申万Ⅰ级和流通市值的行业和市值中性化。

$$factors_i=\beta_i\cdot\ln(MktVal_i)+\sum_{i=1}^{n}\beta_j\cdot Industry_{j,i}+\epsilon_i \tag{2.2}$$

2.2　因子筛选

2.2.1　因子选择

由于公司经营状况能够通过披露的财务报表反映出来，其常用于衡量公司当前的投资价值，对财务报表的分析使得投资者能够从多角度了解企业内在发展状况，知晓公司当期的盈利能力及成长状况等；公司股票的换手率和价格动量等市场行情指标则能够直观得反映出当前股票的市场热度和价格趋势，对股价走势能够提供一定的解释。因此为了反映出公司当前的财务状况和股票市场市场行情对未来收益的影响，本文从公司财务基本面和股票技术指标两个角度选取盈利因子、成长因子、情绪因子、动量因子 4 大类共 40 个具体因子，具体因子及计算方式如表 3.1。其中盈利和成长因子选取公司季度报表数据进行滚动计算获得，情绪和动量因子则由股票日度数据计算获得。

表 2.1　因子表及计算方式

大类因子	具体因子	计算方式
盈利因子	ROA	净利润/总资产
	ROE	净利润/股东权益
	ROIC	归母净利润/投入资本
	OPM	营业利润/营业收入
成长因子	NPG	（当期净利润-上期净利润)/上期净利润
	TAG	(当期总资产-上期总资产)/上期总资产
	ORG	(当期营业收入-上期营业收入)/上期营业收入
杠杠因子	ETOA	股东权益/总资产
	DTOA	总负债/总资产
	CUR	流动资产合计/流动负债合计
估值因子	PB	当前每股市场价格/每股净资产
	PE	当前每股市场价格/每股净利润
	PS	当前每股市场价格/每股销售额
	PCF	当前每股市场价格/每股现金流
情绪因子	VOL_5	5 日换手率均值
	VOL_10	10 日换手率均值
	VOL_20	20 日换手率均值
	TV_20	20 日换手率标准差
	VEMA_10	5 日成交量移动平均
	VSTD_10	10 日成交量标准差
	VSTD_20	20 日成交量标准差
	TRIX_5	MAP=收盘价的 5 日指数平均的 5 日指数平均的 5 日指数平均 $TRIX=(MAP_t-MAP_{t-1})/MAP_{t-1}*100$
	TRIX_10	$TRIX=(MAP_t-MAP_{t-1})/MAP_{t-1}*100$
动量因子	BIAS_5	(收盘价-收盘价的 5 日简单平均)/收盘价的 5 日简单平均
	BIAS_10	(收盘价-收盘价的 10 日简单平均)/收盘价的 10 日简单平均
	BIAS_20	(收盘价-收盘价的 20 日简单平均)/收盘价的 20 日简单平均
	BIAS_60	(收盘价-收盘价的 60 日简单平均)/收盘价的 60 日简单平均
	ROC_6	(收盘价-6 日前收盘价)/6 日前收盘价*100
	ROC_12	(收盘价-12 日前收盘价)/12 日前收盘价*100

续表 1

ROC_20	(收盘价-20 日前收盘价)/20 日前收盘价*100
CCI_10	TYP 典型价格=(最低价+最高价+收盘价）/3 CCI=(TYP-TYP 的 10 日移动平均)/(0.015*TYP 的 10 日平均绝对偏差
CCI_15	CCI=(TYP-TYP 的 15 日移动平均)/(0.015*TYP 的 15 日平均绝对偏差
MAC_5	5 日移动均线/收盘价
MAC_10	10 日移动均线/收盘价
MAC_20	20 日移动均线/收盘价
BOLLD	收盘价的 20 日移动平均-2*收盘价的 20 日标准差
BOLLU	收盘价的 20 日移动平均+2*收盘价的 20 日标准差
PLRC_6	1-6 日收盘价与日期序号(1-6)的线性回归系数
PLRC_12	1-12 日收盘价与日期序号(1-12)的线性回归系数
PLRC_24	1-24 日收盘价与日期序号(1-24)的线性回归系数

2.2.2 因子相关性检验

在因子选取过程中，如 ROA、ROE 等指标来源于公司报表披露的与当期利润和资产相关的财务数据，相关财务基本面因子之间容易出现自相关问题。同时技术面因子计算方式虽有不同，但其基础指标多与收盘价等交易价格数据相关，并且指标时间范围较为接近，也容易出现自相关问题，因此首先通过因子间的相关系数来剔除存在高度相关的因子，因子相关性情况如下图。

在盈利因子中，用来衡量公司回报率的 ROA、ROE 和 ROIC 的指标构成较为相似，容易出现相关问题，拟剔除冗杂的盈利因子。在情绪因子中，基于换手率的 VOL 因子和基于成交量的 VSTD 因子之间也存在一定程度的相关，这主要是由于在股票交易过程中，单一股票往往容易呈现快速上涨或下跌情况，并持续一段时间。在该段交易日中，随着股票上涨加速等情况，股票的换手率和成交量往往呈现一致的变动情况，致使在短期 5 日，10 日的换手率等指标呈现高度相关。为降低模型训练的复杂程度，考虑将相关冗杂因子筛选后剔除。

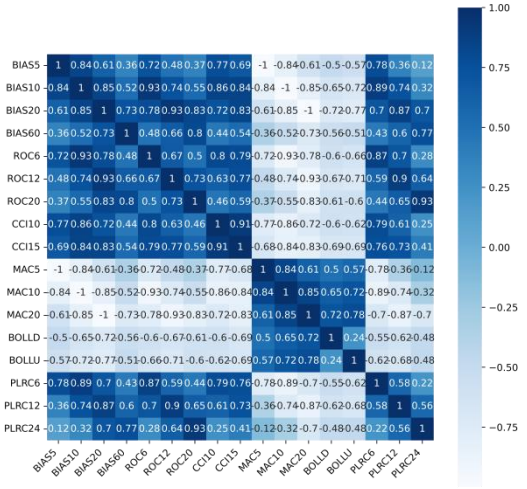

图 2.1　动量因子相关性

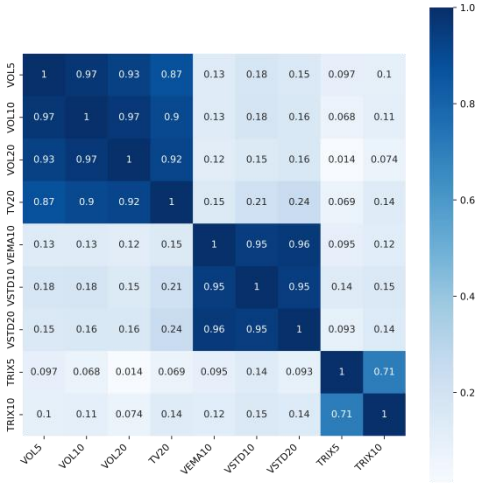

图 2.2　情绪因子相关性

图 2.3　盈利和成长因子相关性

图 2.4　杠杆和估值因子相关性

2.2.3　因子有效性检验

本文依据因子 IC 值评价对因子进行有效性检验。

IC 值即信息系数，其通过因子值与股票下一期收益率的 Pearson 相关系数来表示因子值与下一期股票预期涨跌幅度的相关性强弱。在计算过程中，各因子值为每个调仓日的截面数据，下一期收益率为下一持仓周期的收盘价涨跌幅情况。其中 IC 值区间在-1 至 1 之间，理论上来说 IC 均值的绝对值越大，该因子的预测效果越好，当 IC 值为零时代表该因子与股票收益率无关，为无效因子。IC 值计算方法如下：

$$IC = Corr(F_t, R_{t+n}) \qquad (2.3)$$

其中，F_t 为 t 期的因子值，R_{t+n} 为下一调仓期（n 日）内的股票收益率。

为了避免连续型因子值对有效性分析的影响，本文采用 Rank IC 进行有效性筛选，Rank IC 没有对变量的分布进行假设，对金融数据上有更强的适应性。Rank IC 通过在截

面数据上的因子值排名与下期收益率排名的相关系数来表示因子值与下一期股票预期涨跌幅度的相关性强弱。Rank IC 计算方法如下：

$$Rank\ IC\ =\ Corr(Order_t^F, Order_{t+n}^R) \quad (2.4)$$

其中，$Order_t^F$为 t 期的因子值排名，$Order_{t+n}^R$为下一调仓期（n 日）内的股票收益率排名。

本文测定各因子值 Rank IC 的均值、标准差、信息比率 IR、IC 大于 0 的比率以及 IC 绝对值大于 0.02 的比例。通过以上指标可以评价因子的显著性、稳定性和作用方向稳定性等性质，便于对因子进行比较全面的判断。

综合因子相关性检验和有效性分析的结果，筛选出如表 2.2 的 24 个具体因子作为最终的因子池。

表 2.2　因子 Rank IC 数据表

具体因子	IC 均值	IC 标准差	信息比率 IR	IC>0	|IC|>0.02
ROIC	-0.0233	0.0747	-0.3120	0.3750	0.7083
OPM	-0.0201	0.0610	-0.3295	0.3125	0.7708
NPG	-0.0183	0.0812	-0.2251	0.3750	0.7500
TAG	-0.0241	0.0868	-0.2773	0.4583	0.8125
ORG	-0.0160	0.0900	-0.1780	0.4375	0.8542
PE	-0.0217	0.0968	-0.2242	0.4583	0.8750
PS	-0.0286	0.1071	-0.2673	0.4583	0.7917
PB	-0.0378	0.1378	-0.2745	0.4792	0.8750
VOL5	-0.0461	0.2053	-0.2245	0.4375	1.0000
TV	-0.0435	0.1826	-0.2385	0.4167	0.9375
TRIX10	-0.0357	0.2022	-0.1764	0.4792	0.9375
TRIX5	-0.0212	0.1797	-0.1177	0.5000	0.9167
BIAS5	-0.0254	0.1991	-0.1275	0.4375	0.9375
BIAS20	-0.0271	0.1896	-0.1431	0.4792	0.9375
BIAS60	-0.0562	0.1983	-0.2836	0.3958	0.9375
ROC6	-0.0266	0.1780	-0.1495	0.4375	0.8958
ROC12	-0.0319	0.1820	-0.1751	0.4375	0.9167
CCI10	-0.0293	0.1872	-0.1565	0.4167	0.9375
MAC5	0.0254	0.1992	0.1274	0.5625	0.9583

续表 1

MAC10	0.0255	0.1825	0.1398	0.5625	1.0000
BOLLD	0.0303	0.1742	0.1740	0.5625	0.8958
BOLLU	0.0185	0.2023	0.0914	0.5208	0.9583
PLRC6	-0.0384	0.1757	-0.2187	0.4167	0.9583
PLRC24	-0.0277	0.1979	-0.1399	0.4792	0.9375

可以发现经过因子有效性检验后剩余因子主要为情绪因子和动量因子，且因子 IC 值多为负数，意味着因子值与未来收益率呈现反向变动关系，如股票换手率上升时其未来收益率下降，因子主要呈现为反转效应而不是动量效应。部分研究也表明我国股票市场上日频率的动量可以引发“正反馈”的趋势效应，使得市场反应过度，表现出动量效应；在周频率上则表现出调整快、强度大的反转效应；月频率上则反转强度有所下降，在 A 股市场上月度和周度频率的反转效应仍显著存在。[9]在市场呈现反转效应的情况下，如基本面因子和技术因子都可能会出现与逻辑相悖的情况，这也侧面反应出我国股票市场投资在一定程度上受到投资者情绪的影响，会出现反应过度或反应不足的现象，甚至因股票市场的投机炒作使得财务状况下降甚至不良的公司出现股票价格大幅上涨，而业绩良好的公司却受到资金冷落，没有表现出预期的股价上涨，在多种因素影响下使得股票价格偏离内在价值。

3　多因子选股模型构建与回测

3.1　模型构建

3.1.1　训练及回测环境设置

与传统的机器学习采用的随机拆分和 K 折交叉验证不同，在金融量化研究中面对的多为时间序列数据集，随机拆分等训练集和测试集验证方法容易引入未来函数，导致模型预测效果与测试效果存在较大差别，因此本文采用滚动窗口的方式对因子数据集进行训练，训练集划分方式如图 3.1。

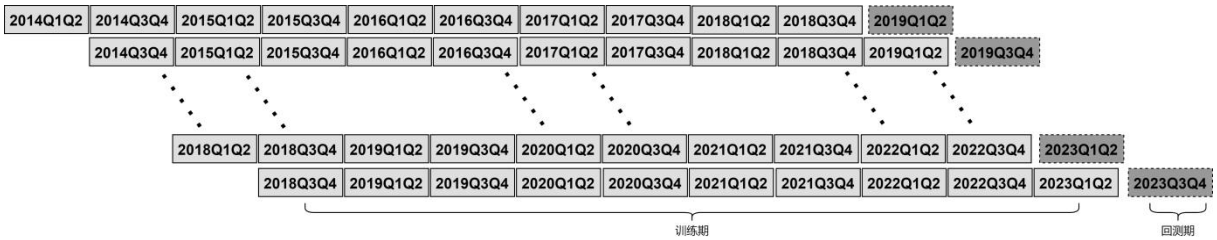

图 3.1　滚动窗口训练示意图

在训练时间上，研究以五年作为一个训练期，预测期为半年。首先将股票未来收益率按照排名分为五组，其中组一代表预测收益最高的 20%，组五代表预测收益最低的 20%。再通过机器学习模型对当日各股票因子数据进行收益率的预测，获取预测收益组标签。

回测基本环境设置如下：

（1）基准：中证 800 指数

（2）滑点：当前价格加减 0.1%

（3）交易费用：卖出印花税千分之一；买入佣金万分之三；卖出佣金万分之三

3.1.2　模型介绍

在获得选股因子的基础上，多因子模型有因子排序、因子打分和因子回归等方法来进行股票筛选，因子排序是通过对首要因子进行排序并选择排名前列的股票，再对次要因子进行排序筛选进而得到最终的选股集合；因子打分是将各股票的各因子排名作为该股票在该因子上的得分，并通过市值加权、因子加权等方式获得各股票的因子总得分，筛选得分较好的股票作为选股集合；因子回归则是通过多元回归来评估各因子与未来收益率之间的线性关系，得到各因子的回归系数作为因子权重，最终得到各股票的未来收

益率。本文则采用随机森林、LightGBM 等机器学习模型对因子数据进行训练拟合，并获得各因子值权重对股票未来收益率进行预测。

随机森林是基于树的集成学习方法，其以决策树作为基分类器，通过训练过程中引入随机性来提高模型的泛化能力和准确性。其核心思想是通过自助法抽样即有放回抽样从数据集中抽取多个不同的子集；在构建每棵树时会从所有特征从所有特征中随机选择一部分特征当作候选特征，再从这些特征中寻找最优的特征进行节点分裂；随机森林的每棵树都可以给出一个分类或回归结果，再通过投票或平均的方法输出最终的结果。由于其引入了随机性并进行树投票来输出结果，其能够在不降低性能的情况下减少模型复杂度，减少模型过拟合问题。

LightGBM 基于梯度提升框架，其核心特点是基于直方图的算法和梯度单边采样，基于直方图算法是将连续数值分散成离散的区间并在每个区间上累计梯度信息，而梯度单边采样是在减少数据量和保持数据代表性的前提下，选择性地忽略了具有小梯度的数据，而保留了具有大梯度的数据，在保持模型准确性的前提下减少计算量。在决策树子模型上，其使用按叶子分裂的方式来分裂节点，在基于 Histogram 的决策树算法下特征值被分为多个部分并在每个部分内寻找分裂。并且其能够特征并行和数据并行，能够有效利用计算资源，提升训练速度。

CatBoost 也基于梯度提升框架，与其他梯度提升树算法类似，通过训练一些列决策树来逐步提升模型的预测性能。其能够克服梯度偏差和预测偏移，通过将样本随机打乱，每个样本只用排序在他前面的样本来训练模型，来估计预测结果的偏度。与 LightGBM 一个个建立节点不同，其使用完全对称二叉树作为基树，特点是在每一层都使用相同的分割特征，这使得模型能够在一定程度上减少过拟合。同时该模型对数据有更大的包容性，其采用的 Ordered Target Statistics 编码方式会对数据集进行多次随机排序，对于每个样本的该类别特征中的某个取值，基于该样本之前的类别编码值取均值，同时加入先验的权重系数，转换为数值型结果。在多次随机排序和迭代下模型的方差会更小，减少模型的过拟合。

TabNet 是基于注意力机制和稀疏特征选择的神经网络模型，通过注意力机制，模型可以自动聚焦于对预测结果更为重要的特征，提高模型对关键信息的感知能力，而稀疏特征选择则是自动选择重要特征而将其他无关特征丢弃，以此降低模型的复杂度和计算成本。该模型还通过残差连接和逐步增强的结构来堆叠多个决策步骤并提取和学习数据特征。

3.1.3 模型超参数优化

机器学习模型的预测性能受到模型自身参数设置的影响，为了提高模型的预测选股性能，本文对选择的随机森林、CatBoost 和 LightGBM 等模型进行超参数优化，参数均通过随机搜索 RandomizedSearchCV 在设定的超参数范围内进行搜索，部分参数如迭代次数等则手动调整，各模型参数设置如表 3.1。

表 3.1 模型超参数优化

模型	超参数	搜索范围	参数设置
随机森林	max_depth	np.linspace(10,500,50)	10
	min_simples_leaf	[1,2,4,8,16]	2
	min_simples_split	[2,5,10]	5
CatBoost	learning_rate	np.logspace(np.log10(0.01),np.log10(0.2),base=10,num=20)	0.05
	depth	range(1,11,1)	9
	random_strength	range(1,21,1)	17
	bagging_temperature	np.linspace(0,1,20)	0.42
TabNet	n_d	range(8,65,8)	16
	n_a	range(8,65,8)	16
	n_steps	range(3,11,1)	3
LightGBM	learning_rate	np.logspace(np.log10(0.01),np.log10(0.2),base=10,num=20)	0.02
	num_leaves	range(2,128,8)	34
	max_depth	range(2,10,1)	6

3.1.4 Stacking 集成

除对比单一机器学习模型的效果外，本文还通过 Stacking 方法将多种分类算法集成起来，在原始股票因子数据的基础上，通过基础学习器对股票未来涨跌幅度进行预测，再以多种预测结果作为新的机器学习特征，对股票涨跌进行预测。

Stacking 方法通常分为训练和融合两个阶段。在训练阶段，通过时序交叉验证的方式对 3 个基学习器进行训练，获得 16 年至 23 年的基学习器的预测结果；在融合阶段，

将各模型的预测结果作为新的特征值输入到元学习器中并获得最终的预测收益组标签。Stacking 集成方法如图 3.2。

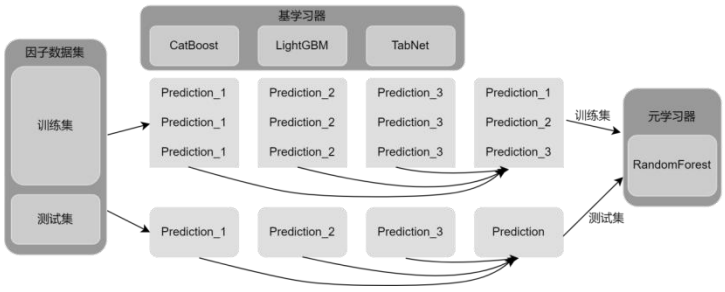

图 3.2　Stacking 集成方法

3.2　模型回测

3.2.1　预测模型对比

为了判断单一机器学习模型和 Stacking 集成模型的预测效果，本文在 2019 年至 2023 年的五年间对 RandomForest、CatBoost 等五个模型进行回测，回测股池为预测标签为 5，即预测收益涨幅最高的组，回测收益走势如图 3.3。

图 3.3　模型回测收益走势

在 2019 年至 2023 年间，A 股市场中证 800 指数经历了一轮上涨行情，至 2021 年 2 月到达顶点，随后开启下跌行情，该期间指数基准收益为 19.47%。四种模型均跑赢指数，其中 LightGBM 表现最好，Stacking 集成表现次之，RandomForest 表现较差，CatBoost 和 TabNet 表现则较为接近。2020 年第一季度前各模型较基准指数没有体现出较好的表现，但在其后一段时间市场整体的下跌和上涨中模型能够筛选出涨势较好的股票，在取得超越基准的收益下控制回撤。LightGBM、CatBoost 和 TabNet 三种模型在 2021 年第三季度前表现较为接近，但在随后的市场下跌行情中 LightGBM 的回撤控制更为优秀，这使得其在整体收益中表现最佳。RandomForest 模型在 2021 年六月前表现明显不如基

准指数，从收益走势上看，该期间其并没有筛选出优势股票，在市场整体大幅上涨时投资组合表现平平。六月后模型收益开始超越基准指数，但回撤风险也超过基准指数。

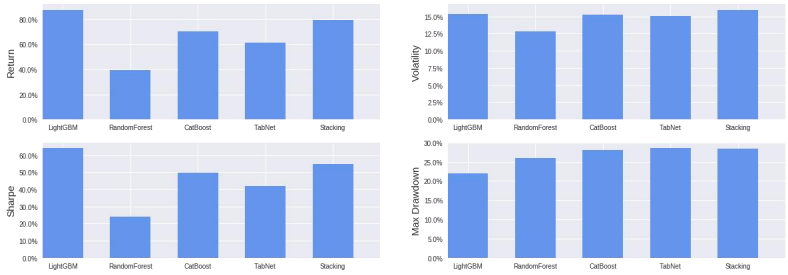

图 3.4　模型评价指标

各指标中，LightGBM 模型在收益和回撤上表现最佳，其在获取最高年化收益 13.86%的同时将回撤控制在 20%左右；其在波动率指标上要略高于其他三种，选股投资组合的风险较大，未来可能会面临更大的回撤等情况，对组合收益率造成影响。各模型指标如表 3.2。

表 3.2 模型评价指标数据表

模型	年化收益率	超额收益率	夏普比率	波动率	最大回撤
RandomForest	7.08%	16.61%	0.239	0.129	26.05%
LightGBM	13.86%	57.05%	0.642	0.154	21.95%
CatBoost	11.58%	42.40%	0.496	0.153	28.05%
TabNet	10.33%	34.82%	0.419	0.151	28.57%
Stacking	12.80%	50.07%	0.550	0.160	28.36%

3.2.2　参数分析

在证券投资中，诸如投资者的股票投资期和选股调仓周期等参数都会对选股投资组合的收益造成影响，本小节将针对这些参数，使用 LightGBM 和 Stacking 集成模型进行模型的稳健性分析。

在进行调仓周期调整时，回测资金量为 1000 万元。可以发现在不同的持仓周期设定下，都能够取得一定的超额收益。在 10 天的周期下，策略会进行频繁的调仓，但调仓产生的收益未必大于因频繁买入卖出操作产生的交易成本，致使策略收益下降。在两种模型中设定 50 天调仓的收益和回撤控制都表现最好，综合看 LightGBM 表现更为稳定。不同参数下投资收益情况如表 3.3。

表 3.3　调仓周期设定收益表

		LightGBM				Stacking 集成			
		年化收益	最大回撤	费用（万元）	夏普比率	年化收益	最大回撤	费用（万元）	夏普比率
持仓周期	10 天	8.66%	24.33%	121.10	0.301	8.96%	28.11%	110.37	0.300
	20 天	10.32%	23.87%	68.86	0.403	11.44%	28.17%	63.95	0.447
	30 天	13.86%	21.95%	50.95	0.642	12.80%	28.36%	44.81	0.550
	40 天	10.40%	25.64%	35.70	0.412	11.10%	29.42%	32.81	0.434
	50 天	15.53%	21.53%	33.97	0.747	18.15%	24.57%	32.52	0.847
	60 天	14.96%	25.15%	29.76	0.701	14.17%	30.64%	26.36	0.625

参考当前的量化主动基金发展情况，至 2023 年基金规模有半数集中于 1 亿元以下，仅有少数基金规模在 10 亿元以上，因此本文将机构量化投资者的资金量划分为 5000 万、1 亿元和 5 亿元，而针对个人投资者则设定资金量为 50 万、100 万、500 万和 1000 万元。设定回测资金量为以上数据，测定不同资金量情况下策略收益情况。

在其他参数相同的情况下仅调整投资总金额，可以发现当资金量达到 500 万元后收益开始趋于稳定，各项指标基本保持一致。由于股票 100 股起买的限制，在小资金量情况下使用等权重分配资金，当选股标的股票较多时，单个股票的可用资金较少，仅能成功购买股池中的部分低价股。在 50 万元的资金量下年化收益有所下降，但同时持股数的减少也使得最大回撤得到了很好控制。

表 3.4　资金量设定收益

		LightGBM			Stacking 集成		
		年化收益	最大回撤	夏普比率	年化收益	最大回撤	夏普比率
资金量	50 万	8.51%	10.39%	0.564	6.91%	13.01%	0.383
	100 万	11.17%	15.08%	0.625	9.57%	21.16%	0.478
	500 万	13.73%	21.37%	0.647	12.46%	27.83%	0.540
	1000 万	13.86%	21.95%	0.642	12.80%	28.36%	0.550
	5000 万	13.98%	22.33%	0.639	13.04%	28.76%	0.555
	1 亿元	14.01%	22.37%	0.639	13.06%	28.80%	0.556
	5 亿元	14.02%	22.40%	0.639	13.04%	28.76%	0.555

3.2.3 投资组合优化

原策略采用等权重的方法来配置选股股票，该方法下资金被平均分配给各支股票，风险也被分散其中。由于该方法不需要考虑各股票资产间的收益和风险情况，故使用门槛较低，但不同的股票各有各的特点，等权重配置下也容易错失某些优质股票的投资机会。本节在原有的选股策略的基础上，针对调仓日待买入股票进行最小方差组合、风险平价的资金权重优化，以期在获得更高收益的同时降低风险。

最小方差组合方法以 Markowitz 的均值-方差模型为基础，在均值-方差模型中，有效边界及其右侧的点构成投资可行集，在有效边界上的点存在期望收益率一定下组合风险最小和风险一定下期望收益率最高两种情形。考虑组合权重为$w^T = (w_1, w_2, \cdots, w_N)$，对应收益率为$r^T = (u_1, u_2, \cdots, u_N)$，各资产收益率协方差矩阵为$\Sigma$，则投资组合的预期收益率为$r_p = w^T r$，组合方差为$\sigma_p^2 = w^T \Sigma w$。通过求解最小方差或最大收益得到组合的权重分配：

$$\begin{cases} min \quad w^T \Sigma w \\ s.t.\ w^T r = r_p \end{cases} \quad 或 \quad \begin{cases} max \quad w^T r \\ s.t.\ w^T \Sigma w = \sigma_p^2 \end{cases} \tag{3.1}$$

通过最大化效用函数：

$$\mathrm{U} = \mathrm{r} - \frac{\delta}{2}\sigma_{\mathrm{p}} = \mathrm{w}^{\mathrm{T}}\mathrm{r} - \frac{\delta}{2}\mathrm{w}^{\mathrm{T}}\Sigma\mathrm{w} \tag{3.2}$$

令 U 对 w 的一阶导数为 0 可以求解出目标资产配置权重：

$$w^* = (\delta\Sigma w)^{-1}\mathrm{r}. \tag{3.3}$$

最小方差方法即求解出风险最小的资产配置组合，组合通过以下关系求解：

$$w^* = arg\ min\frac{1}{2}w^T\Sigma\mathrm{w} \tag{3.4}$$

$$\begin{cases} w^T t = 1 \\ w_i \geq 0 \end{cases}. \tag{3.5}$$

但由于该方法基于均值方差模型，其同样受到有效市场、理性投资者和风险厌恶等假设的限制，并且通过最小方差分配权重可能会出现中高风险资产权重为 0 的极端情况，使得投资股票数量减少，甚至集中于个别股票上。

风险平价方法通过平衡不同资产在组合风险中的贡献程度来实现投资组合风险结构的优化，在该方法中投资组合不会暴露在单一资产的风险敞口中，可以获得较为稳健的投资组合。风险平价方法权重求解如下：

$$\sigma_p = \sqrt{\sum_{i=1}^{N} w_i\sigma_i^2 + 2\sum_{i=1}^{N}\sum_{j>1}^{N} w_i w_j \sigma_{ij}} \tag{3.6}$$

$$RC_i = w_i\frac{\partial_{\sigma_p}}{\partial_{w_i}} = \frac{w_i^2\sigma_i^2 + \sum_{i=1}^{N} w_i w_j \sigma_{ij}}{\sigma_p} \tag{3.7}$$

其中，σ_i为资产 i 的方差， σ_p为组合风险，w_i为各资产在组合中的权重。

当每类资产 RC 相等时资产的组合风险贡献程度相等，通过求解如下最优解即可得到风险平价方法下各资产的配置权重：

$$\begin{cases} min\sum_{i=1}^{N}\sum_{j=1}^{N}\left(RC_i - RC_j\right)^2 \\ \sum_{i=1}^{N} w_i = 1, 0 \le w_i \le 1 \end{cases} \tag{3.8}$$

通过使用最小方差方法和风险平价法对投资组合进行优化，其中股票历史收益率等数据的时间区间为回测时点前 250 个交易日，得到回测收益曲线如图 3.5。可以发现在等权、最小方差和风险平价三种方法下投资组合的五年收益较为接近，其中风险平价方法下的投资收益略低，可能是对资产的风险估计不准确等原因导致。而在最小方差方法下，投资组合的收益在市场整体面大幅下跌风险时可以控制投资组合风险，在市场下行中也表现出较好的收益稳定性。

图 3.5　投资组合优化收益

在对投资组合权重的参数中，由于没有对单支股票的最低权重和最高权重进行限制，在每个调仓日进行的最小方差优化中容易出现极端的权重偏离，少数股票占据了最大的权重。因此在限定单一股票最大权重为 5%~50%的范围内进行测试。测定的各组合投资收益如表 3.5。

表 3.5　最小方差组合权重优化

限制权重	年化收益	最大回撤	夏普比率
5%	13.91%	15.55%	0.75
10%	13.24%	13.94%	0.715
20%	13.47%	14.54%	0.749
30%	14.71%	14.29%	0.849
40%	14.80%	14.31%	0.854
50%	14.28%	14.76%	0.811
无限制	10.25%	17.04%	0.494

4 情绪择时

机器学习模型在各类风格因子基础上对股票未来收益率有较好的预测效果，但投资策略最终是否有好的投资表现不仅仅与选股集合有关，诸如市场择时、仓位控制等都是影响策略收益的重要因素。上一章基于 LightGBM 模型构建出一个能够跑赢市场指数，获取超额收益的多因子选股模型，并且通过对比分析确定了投资环境的各个参数以及资产组合配置方法。但前文仅立足于个股的因子表现方面，通过机器学习对未来收益率进行预测，缺乏对市场行情的整体判断，在市场下行行情中开仓投资，尽管收益跑赢基准指数，仍可能会面临组合资产亏损的风险。在策略采取全仓位投资的情况下，能否通过对市场整体行情的判断来控制策略的开仓与否则显得尤为重要。

在市场行情改变过程中，投资者的情绪也会随之有较大变化，在行为金融学中部分研究已经表明投资者行为在一定程度上会影响到股票的价格，而当市场行情处于极端变化中，大部分投资者都会采取相近的行为来规避行情变化带来的投资风险。如市场大幅下跌情况下，投资者会倾向于抛售持有的资产，而潜在的投资者则会暂缓入市，等待市场行情的好转；政府会暂缓企业 IPO 以避免新上市公司对二级市场的“抽血”，避免市场进一步下跌，预期上市的公司也可能会选择暂缓上市以期未来能够获得更高的企业估值。投资情绪可以用来反映当前及预期未来市场的整体情况，为策略提供一种有效的指标来判断市场是否适合策略开仓投资，因此，本节采用主成分分析法构建情绪择时指标，通过反映投资者情绪变化对策略的开仓、空仓情况进行控制。

4.1 情绪指标的构建

4.1.1 情绪变量选取

本文通过主成分分析的方法构建情绪指数，为了反映投资者行为以及市场宏观环境对股票价格的影响，本节选取指数换手率、新增开户人数、银行同业拆借利率、M1 货币供应量、基金折价率和 IPO 个数六个指标作为情绪变量。

指数换手率从微观层面反映当前市场的热度，当换手率上升时意味着投资者的交易情绪高涨，其可能意味着股市即将开启上涨行情。当换手率高于正常值时则市场可能已经超涨，证券价值已经极大偏离其内在价值，市场处于狂热的投机情绪当中，后市可能会存在大幅下跌的行情。基金折价率反映出一段时间内封闭式基金投资者的交易热情，当折价率走低，交易价格上升时往往认为投资者对后市看涨，而但市场上涨一段时间后折价率的上升又表明投资者对当前股票市场投资情绪的保守倾向。

M1 货币供应量和银行同业拆借利率则从宏观层面反映当前经济运行情况。广义货币供应量 M1 反映出居民和企业的资金松紧变化，当 M1 增速加快时意味着资金较为活跃。当股市赚钱效应较为明显时，部分投资者会选择将资金投入股市，增大 M1 货币量。银行拆借利率是短期利率的风向标，其能够反映出货币政策走向，反映出市场的流动性松紧程度，进而反映出股票市场收益情况。

IPO 个数和开户人数则从市场和投资者两方面反映股票市场情绪。当股市行情向好时，企业 IPO 更容易获得更高的估值，而在下行行情中则容易出现“破发”的情况，行情较差时企业的 IPO 动力不足。当市场处于上涨的投资情绪中容易吸引民众参与股票投资，开户人数也会随之上升，而当市场遇冷时则投资者没有投资动力，也没有开户投资的动力。

本文获取 2015 年 6 月至 2023 年 8 月共 99 个月的情绪变量数据，频率为月度，数据来源于国泰安数据库。

4.1.2 主成分分析

使用 PCA 需要变量间存在较强的线性关系，本节首先对选择的各个情绪变量进行 KMO 和 Bartlett 球形检验，一般情况下需要 KMO 值大于 0.5 并且 Bartlett 检验显著，结果如表 4.1。

表 4.1 KMO 和 Bartlett 检验

KMO 值		.662
Bartlett 球形度检验	近似卡方	1587.158
	自由度	66
	显著性	.000

可以发现 12 个指标的 KMO 值大于 0.6，并且 Bartlett 检验中 P 值小于 0.01，在 1%置信区间上拒绝原假设，可以进行主成分分析。总方差解释如表 4.2。

表 4.2 总方差解释

成分	初始特征值		
	特征根	方差百分比	累积 %
1	4.299	35.824	35.824
2	3.159	26.325	62.149
3	1.643	13.688	75.838
4	1.119	9.324	85.161

续表 1

5	0.614	5.113	90.274
6	0.420	3.497	93.771
7	0.306	2.553	96.324
8	0.244	2.031	98.355
9	0.129	1.074	99.429
10	0.044	0.369	99.797
11	0.024	0.197	99.995
12	0.001	0.005	100.000

可以看出前 4 个成分的特征值大于 1，为了有较强的解释力度，本文提取前 4 个特征值大于 1 的成分，其累计方差解释率达到 85%以上。第一阶段主成分分析得到的成分矩阵如表 4.3。

表 4.3 成分矩阵

情绪变量	成分			
	1	2	3	4
TURN	0.510	0.640	-0.403	0.055
NA	0.366	0.649	-0.024	0.292
IBOR	-0.745	0.337	0.468	0.175
MS	0.625	-0.701	-0.103	0.168
DCEF	0.675	0.246	0.442	-0.511
IPON	0.614	-0.122	0.492	0.387
$TURN_{t-1}$	0.467	0.654	-0.411	0.016
Na_{t-1}	0.380	0.726	-0.055	0.309
$IBOR_{t-1}$	-0.777	0.332	0.389	0.185
MS_{t-1}	0.623	-0.704	-0.103	0.171
$DCEF_{t-1}$	0.644	0.290	0.461	-0.501
$IPON_{t-1}$	0.598	-0.061	0.516	0.389

获取成分矩阵后通过如下公式计算情绪指数 EI：

$$EI = \sum_{i=1}^{t=4} w_i \times p_i \div \sqrt{\mu_i} \quad (4.1)$$

其中，w_i为成分 i 方差百分比/前 t 个成分累计百分比，p_i为成分 i 的成分值，μ_i为成分 i 的特征值。

得到情绪指数 EI 的权重计算公式：

$$EI = 0.111 \cdot TURN + 0.14 \cdot NA - 0.01 \cdot IBOR + 0.006 \cdot MS + 0.119 \cdot DCFE + 0.134 \cdot IPON + 0.103 \cdot TURN_{t-1} + 0.149 \cdot NA_{t-1} - 0.021 \cdot IBOR_{t-1} + 0.006 \cdot MS_{t-1} + 0.122 \cdot DCFE_{t-1} + 0.141 \cdot IPON_{t-1} \quad (4.2)$$

4.2 因果检验与回测

4.2.1 情绪指标与大盘收益

为判定情绪指数与大盘走势的关系，为情绪择时指标的区间范围的选择提供依据，本文采用格兰杰因果检验来分析情绪指数和上证指数下月收盘价之间的相关关系。首先进行平稳性检验，检验结果如表 4.4。

表 4.4 ADF 检验

变量	类型	ADF 值	1%水平	5%水平	10%水平	P 值	结果
EI	C,T,0	-1.811	-4.063	-3.461	-3.156	0.691	不平稳
D(EI)	C,T,0	-11.411	-4.058	-3.458	-3.155	-0.000	平稳
$Close_{t+1}$	C,T,0	-3.747	-4.055	-3.457	-3.154	0.024	不平稳
$D(Close_{t+1})$	C,T,0	-10.432	-4.056	-3.457	-3.155	0.000	平稳

检验结果发现原序列都存在单位根，为非平稳序列。在对原序列进行一阶差分后为平稳序列，情绪指数和收盘价序列均在同阶滞后情况下通过 ADF 检验。滞后阶数 p 的选择通过多准则联合确定的方法，确定结果如表 4.5，选择阶数为 4 阶进行格兰杰因果检验，检验结果如表 4.6。

表 4.5 VAR 滞后阶数选择结果

Lag	LogL	LR	FPE	AIC	SC	HQ
0	-600.6903	NA	1679.43	13.10196	13.15678	13.12409
1	-595.4568	10.12568	1635.067	13.07515	13.23961	13.14153
2	-571.7297	44.87506	1065.03	12.6463	12.92041	12.75693
3	-563.844	14.57144	979.1269	12.56183	12.94558	12.71671
4	-553.2446	19.12506*	848.8322*	12.41836*	12.91175*	12.61750*
5	-551.488	3.093012	892.1873	12.46713	13.07017	12.71052

表 4.6 格兰杰因果检验结果

Dependent variable: D(CLOSE)				Dependent variable: D(EI)			
Excluded	Chi-sq	df	Prob.	Excluded	Chi-sq	df	Prob.
D(EI)	16.00538	4	0.003	D(CLOSE)	20.36091	4	0.0004
All	16.00538	4	0.003	All	20.36091	4	0.0004

可以发现情绪指数构成对下月收盘价的因果关系，本月的投资者情绪会影响到下月大盘指数收盘价。在此关系上本文通过情绪指数与情绪指数的 6 期移动平均之间的相对关系来判断开仓及空仓时机，当情绪指数大于 6 期移动平均时则开仓，当小于 6 期移动平均时则卖出保持空仓，情绪指数及均线对比如图 4.1。

图 4.1　情绪指数及 6 期移动平均走势

4.2.2　情绪指标回测

加入情绪择时的交易策略年化收益 11.86%，最大回撤 13.71%，相交原策略年化收益下降 2%，最大回撤下降 8.24%。由于情绪择时采用的是月度数据，单个空仓期为一个月，在股市走势中短期的下跌和上涨的时间经常小于一个月，策略难以抓住下跌过程中的短期反弹，而采取空仓，因此情绪择时策略的收益没有提高，但也正因此使得策略可以规避市场下跌风险，策略回撤表现有了较大的提高。加入情绪择时指标的回测收益情况如图 4.2。

图 4.2　情绪择时策略收益走势

5 结论与展望

5.1 研究结论

本文选择中证 800 指数成分股构建基础股票池，通过滚动回测的方法对 2019 年至 2023 年 5 年间的股票市场进行策略回测。其中，在财务基本面和技术基本面基础上通过因子相关性分析和因子 Rank IC 分析筛选了 24 个具体因子进行模型选股。在多因子选股模型算法上则通过对随机树林、LightGBM、CatBooat、TabNet 以及基于 Stacking 集成方法的集成模型的对比分析，筛选出具有较好预测效果的 LightGBM 模型作为因子选股模型，并比较资金量等不同参数下策略的收益表现。在完成基础的模型对比后本文通过最小方差方法和风险平价方法来优化投资组合的分配权重，并通过主成分分析法引入情绪指数，反映当前市场投资的投资情绪。在以该指数的基础上建立情绪择时策略，提升原策略的收益效果。

回测结果表明，LightGBM 在其他机器学习模型中有着较为优秀的收益表现，LightGBM 可以在 5 年的回测期内取得 13.86%年化收益的同时将策略回撤控制在 20%左右，收益远远跑赢大盘，证明本文多因子选股模型的有效性；加入最小方差权重优化有助于降低组合的整体风险，控制策略回撤；加入情绪指数择时的交易策略可以保持较为准确的市场情绪判断效果，通过保持空仓有效规避市场下跌风险，使得策略回撤进一步下降。

5.2 不足与展望

本文基于机器学习模型构建的多因子选股交易策略，通过挖掘过往数据中的信息来预测未来股票的收益情况，虽然取得较好的策略收益，但仍然存在不足有待解决。

第一，在因子选取阶段，由于量化交易策略的广泛使用，相关因子有效性受到市场交易的影响可能不再有效，如出现因子半衰期等现象。本文因子选取选取过程中局限于财务基本面以及简单的技术面因子，在大量挖掘因子的今天，选取的因子在有效性上可能不如其他因子，且因子有效性上没有进行严格的单因子检验，可能会引入部分失效因子，影响模型预测的准确性。

第二，本文在选股范围上基于中证 800 指数的成分股，这在一定程度上使得策略被动关注在大中市值的股票上，而当面对小市值行情的时候，该策略收益可能不尽如人意，需要扩大策略的选股范围，使得策略有更好的市场适应性。

第三，本文策略关注于证券资产，而忽视了对债券、期货等资产的配置，在股票下行行情中，股票很难带来大的收益甚至面临亏损风险。可以通过资产配置方法，将资金分散在债券等不同的大类资产，降低资金整体风险。

第四，本文选取的机器学习模型多为近些年较多使用的 Boosting 梯度提升树模型，其在选股策略中有一定的效果。但使用诸如循环神经网络 RNN 和长短时记忆网络 LSTM 等深度学习模型是否会获得更好的投资效果还需要进一步的实验验证。

参考文献

[1] Basak S, Kar S, Saha S, et al. Predicting the direction of stock market prices using tree-based classifiers[J]. North American Journal of Economics and Finance, 2019, 47: 552-567.

[2] Chatzis S P, Siakoulis V, Petropoulos A, et al. Forecasting stock market crisis events using deep and statistical machine learning techniques[J]. Expert Systems with Applications, 2018, 112: 353-371.

[3] Nabipour M, Nayyeri P, Jabani H, et al. Deep Learning for Stock Market Prediction[J]. Entropy, 2020, 22(8).

[4] Nabipour M, Nayyeri P, Jabani H, et al. Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis[J]. Ieee Access, 2020, 8: 150199-150212.

[5] Wang Y, Guo Y K. Forecasting Method of Stock Market Volatility in Time Series Data Based on Mixed Model of ARIMA and XGBoost[J]. China Communications, 2020, 17(3): 205-221.

[6] Daniel K, Hirshleifer D, Sun L. Short- and Long-Horizon Behavioral Factors[J]. Review of Financial Studies, 2020, 33(4): 1673-1736.

[7] Kim S. Enhanced factor investing in the Korean stock market[J]. Pacific-Basin Finance Journal, 2021, 67.

[8] Stambaugh R F, Yuan Y. Mispricing Factors[J]. Review of Financial Studies, 2017, 30(4): 1270-1315.

[9] 阎畅,江雪. 动量与反转效应在中国股票市场的实证研究——基于时间频率和市场状态的分析 [J]. 投资研究, 2018, 37 (02): 74-86.

[10]袁亦方. 基于机器学习的多因子选股组合预测与 Alpha 对冲策略研究[D].中央财经大学,2022.DOI:10.27665/d.cnki.gzcju.2022.000851.

[11]罗泽南.基于集成树模型的 Stacking 量化选股策略研究[J].中国物价,2021,(02):81-84.

[12]王一卓.基于 Boosting 算法的多因子量化选股实证研究[D].山东大学,2020.DOI:10.27272/d.cnki.gshdu.2020.001615.

[13]王云凯,蓝金辉.ML-FFA:基于机器学习和基本面因子分析的量化投资策略[J].时代金融,2018,(32):358-359+375.

[14]李佩琛.用 Stacking 算法堆积随机森林、GBDT、SVM、Adaboost 等七种算法的多因子选股模型[D].浙江工商大学,2018.

[15]李斌,邵新月,李玥阳.机器学习驱动的基本面量化投资研究[J].中国工业经济,2019,(08):61-79.DOI:10.19581/j.cnki.ciejournal.2019.08.004.

[16]王春丽,刘光,王齐.多因子量化选股模型与择时策略[J].东北财经大学学报,2018(05):81-87.DOI:10.19653/j.cnki.dbcjdxxb.2018.05.011.

[17]黄秋丽,黄柱兴,杨燕.基于递归特征消除和 Stacking 集成学习的股票预测实证研究[J].南宁师范大学学报(自然科学版),2021,38(03):37-43.DOI:10.16601/j.cnki.issn2096-7330.2021.03.008.

[18]刘向丽,崔文泓.一种基于决策树的用于构建量化策略的分类器[J].纯粹数学与应用数学,2023,39(03):339-349.

[19]方毅,陈煜之,卫剑.人工智能与中国股票市场——基于机器学习预测的投资组合量化研究[J].工业技术经济,2022,41(08):83-91.